ABSTRACT

Title of thesis: A STUDY OF FEATURE SETS FOR EMOTION RECOGNITION FROM SPEECH SIGNALS

Yi-Chun Ko, Master of Science, 2015

Thesis directed by: Professor Carol Espy-Wilson
Department of Electrical and Computer Engineering

This thesis focuses on finding useful features for emotion recognition from speech signals. In comparison to the popular openSMILE **'emobase'** feature set, our proposed method reduced the size of feature space to about 28% yet boosted the recognition rate by 3.3%.

Given we are at a point technologically where computing is cheap and fast, and lots of data are available, the approach to solving all sorts of problems is based on sophisticated machine learning techniques to implicitly make sense of data. Yet in this work, we study particular features that are felt to correlate with changes in emotion but have not been commonly selected for emotion recognition tasks. Jitter, shimmer, breathiness, and speaking rate are analyzed and are found to systematically change as a function of emotion.

We not only explore these additional acoustic features that help improve the classification performance, but also try to understand the importance of the existing features in improving accuracy. Our results show that using our features together with MFCCs and pitch related features lead to a better performance.

A STUDY ON FEATURE SETS FOR EMOTION RECOGNITION FROM SPEECH SIGNALS

by

Yi-Chun Ko

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Master of Science
2015

Advisory Committee:
Professor Carol Espy-Wilson, Chair/Advisor
Professor William Idsardi
Professor Jonathan Simon

Acknowledgments

I owe my gratitude to all the people who have helped me through my graduate study years and made this thesis possible.

Firstly, I would like to express my sincere gratitude to my advisor Prof. Carol Espy-Wilson, for her support of my Master's study and related research. Her guidance helped me in all the time of research and writing of this thesis.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. William Idsardi and Prof. Jonathan Simon, for their insightful comments and inspiration.

I thank Dr. Romel Gomez, for his encouragement and mental support during my difficult times.

I thank my fellow labmates Yanbo Xu, Saurabh Sahu, and Ganesh Sivaraman, for the wonderful ideas they shared and the fruitful discussions we had.

Special thanks go to Pin-Hui Chen, Eddie Tseng, Zhung-Han Wu, and Hsueh-Chien Cheng, for their care and for all the good times we had together.

Last but not the least, I would like to thank my family: my parents, Chih-Hung Ko and Chiao-Ling Chiang, and to my brother, Meng-Ting Ko, and sister, Szu-Yu Ko, for supporting me spiritually throughout my years in the States and my life in general. I also want to thank my boyfriend Chia-Chu Chou for his continuous support, encouragement, and his faith in me. I would not have completed my studies without them.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AMDF	average magnitude difference function
APP	aperiodicity, periodicity, and pitch
LLD	low-level descriptors
MFCC	Mel-frequency cepstral coefficients
openSMILE	open-Source Media Interpretation by Large feature-space Extraction
P2FA	Penn Phonetics Lab Forced Aligner

Chapter 1: Introduction

1.1 Motivation

Emotions undoubtedly play an important role in our daily life. Hence the need and importance of automatic emotion recognition has grown with the increasing role of human computer interface applications. Studies that focus on user emotions while he or she interacts with computers and applications belong to the domain of affective computing. It is an interdisciplinary field spanning engineering, computer science, psychology, and cognitive science.

Having a computing device which has the ability to detect and appropriately respond to its user's emotions could be beneficial to many fields. Automatic emotion recognition is a crucial technology that can be used in modern Human-Computer interfaces and has numerous applications as described in [30]. Call center managers would be able to monitor the quality of the services provided by their agents, and handle very angry customers by specially trained agents. In e-learning situations, the computer could detect when the user is having difficulty and offer expanded explanations or additional information. What's more, user's interest, stress, and cognitive load can be employed to adapt the teaching pace in an online tutoring system. In the field of media retrieval, highlights in sports games can be extracted

by measuring the level of excitement of the reporter. Another field of applications is Robotics: With a better modelling of these states and traits, we will be able to add social competence to humanoid or other highly interactive and communicative robots, assistive robots, or to virtual agents.

As a result, a challenging problem of automatic recognition of human emotions has become a research field of large interests. Several information sources such as facial expressions, voice signals, and physiological measurements, can be used for human emotion recognition. Among all, speech signals are the most suitable way for the purpose since it is non-invasive and easier to acquire. It is known that emotions cause mental and physiological changes which are reflected in uttered speech; however, how the speech is affected by emotions is not yet well understood.

1.2 State of the Art Feature Sets

After decades of research, there are no standard agreed-upon acoustic features for the automatic recognition of human emotions. This is because only few data exists. Most of the emotional speech datasets are private and only three are freely available to us. The Berlin emotional database [3] is in German, and it contains about 500 utterances spoken by actors. The LDC Emotional Prosody Speech and Transcripts database [20] is in English and has around 1000 utterances, but the utterances—date and time—are all short. The Electromagnetic Articulography database [18], the one we used, contains around 600 utterances in English produced by three speakers. In addition, there is a large range and set of emotions, requiring

different sets of acoustic features for discrimination. Furthermore, most studies are carried out on single data set, where the types of emotions, the speakers, and the acoustic conditions are the same throughout. Nevertheless, a number of acoustic features have been commonly employed for automatically classifying emotions. These include: prosodic features (pitch, intensity, duration, rhythm) and spectral features (Mel-Frequency Cepstral Coefficients (MFCCs) and alike, formants, spectral statistics). Banse *et al.*(1996) [1] examined vocal cues for 14 emotion categories. The speech features they used are related to the fundamental frequency F0, the energy, the articulation rate, and the spectral information in voiced and unvoiced portions. Schuller *et al.*(2004) [28] ranked more than 200 features with aid of a Linear Discriminant Analysis, and provided a list of their top 33 features in detail. Pitch-related features rank the top in their study. In addition, it is believed that the emotional content of an utterance is strongly related to its voice quality. Li *et al.*(2007) [19] developed their recognition system using continuous HMM as a classifier and applied to utterances from the SUSAS database with the following selected speaking styles: angry, fast, Lombard, question, slow and soft. The baseline accuracy corresponding to using only MFCC as features was 65.5%. The classification accuracy was 68.1% when the MFCC was combined with the jitter, 68.5% when the MFCC was combined with the shimmer, and 69.1% when the MFCC was combined with both of them. A summary of the commonly used features is shown in Table 1.1.

Few years ago, the Munich open Speech and Music Interpretation by Large Space Extraction (openSMILE) tool was developed and it enables us to extract large audio feature spaces in real-time. What's more, it provides different configurations

prosodic	spectral	voice quality
pitch	MFCCs	jitter
intensity	LPCCs	shimmer
duration	formants	
rhythm		

Table 1.1: Commonly used features for automatically classifying emotions

for various purposes, such as emotion recognition, speech recognition, and chord recognition. It quickly became the standard feature-extraction tool for the annual INTERSPEECH Computational Paralinguistics Challenges.

1.3 Objectives

As discussed previously, openSMILE toolkit is widely used in the domain. Based on the emotion recognition features sets (the **'emobase'** configuration) provided by openSMILE, there are two objectives for this work: One is to explore additional acoustic features that improve the performance of emotion recognition from spoken audio signals. The other is to reduce the number of commonly used features for emotion recognition, finding the effective ones. In fact, the **'emobase'** configuration consists of almost a thousand of features.

1.4 Organization of this Thesis

The rest of the thesis is organized as follows. In Chapter 2, we cover the background knowledge of speech production, voice quality and emotions. We also introduce the Penn Phonetics Lab Forced Aligner (P2FA) toolkit [38] and the aperiodicity, periodicity, and pitch (APP) system [9] that help us with extracting features.

In Chapter 3, we outline our findings with respect to various speech parameters. The relationship between speech signals and their intrinsic acoustic features, which could be helpful for decoding the emotion states, will be introduced.

In Chapter 4, we describe the database used and methodology chosen in this work. Afterward, experiments we conducted and corresponding results are presented.

Finally, concluding remarks and suggestions for future works are given in Chaper 5.

Chapter 2: Background

2.1 Speech Production

As shown in Figure 2.1(a), the mechanism of speech production starts from our lungs. In nearly all speech sounds, the basic source of power is the respiratory system pushing air out of the lungs. Air flows from the lungs up to the trachea, into the larynx, and then it passes between the vocal folds. By the time it passes through the vocal folds, if the vocal folds are apart, the air from the lungs will have a relatively free passage into the pharynx and then the nasal or oral cavity. In the case of the latter, the velum is in charge of such selection. But if the vocal folds are adjusted so that there is only a narrow passage between them, the airstream from the lungs will set them vibrating. Sounds produced when the vocal folds are vibrating are said to be voiced, as opposed to those in which the vocal folds are apart, which are said to be voiceless or unvoiced.

The articulation process takes place in the mouth and it is the process through which we can differentiate most speech sounds. In the mouth we can distinguish between the oral cavity, which acts as a resonator, where the frequencies and amplitudes of the resonance are determined by the position of and the articulators: upper and lower lips, upper and lower teeth, tongue (tip, blade, front, back) and roof of

the mouth (alveolar ridge, palate and velum). So, speech sounds are distinguished from one another in terms of where (place) and how (manner) they are articulated.

2.2 The Penn Phonetics Lab Forced Aligner toolkit

The Penn Phonetics Lab Forced Aligner (P2FA) toolkit is developed by Yuan *et al.* [38]. It is a Python script which takes a .wav file and a .txt file orthographic transcript and generates a time-aligned phonetic transcripts of the speech waveform. The toolkit runs based on the Hidden Markov Model Speech Recognition Toolkit(HTK). P2FA uses HTK, the CMU Pronouncing Dictionary, and a set of acoustic models derived from a corpus of recordings of the U.S. Supreme Court. It is extremely useful for scripted databases. The simplest way to use it is to type in the following command:

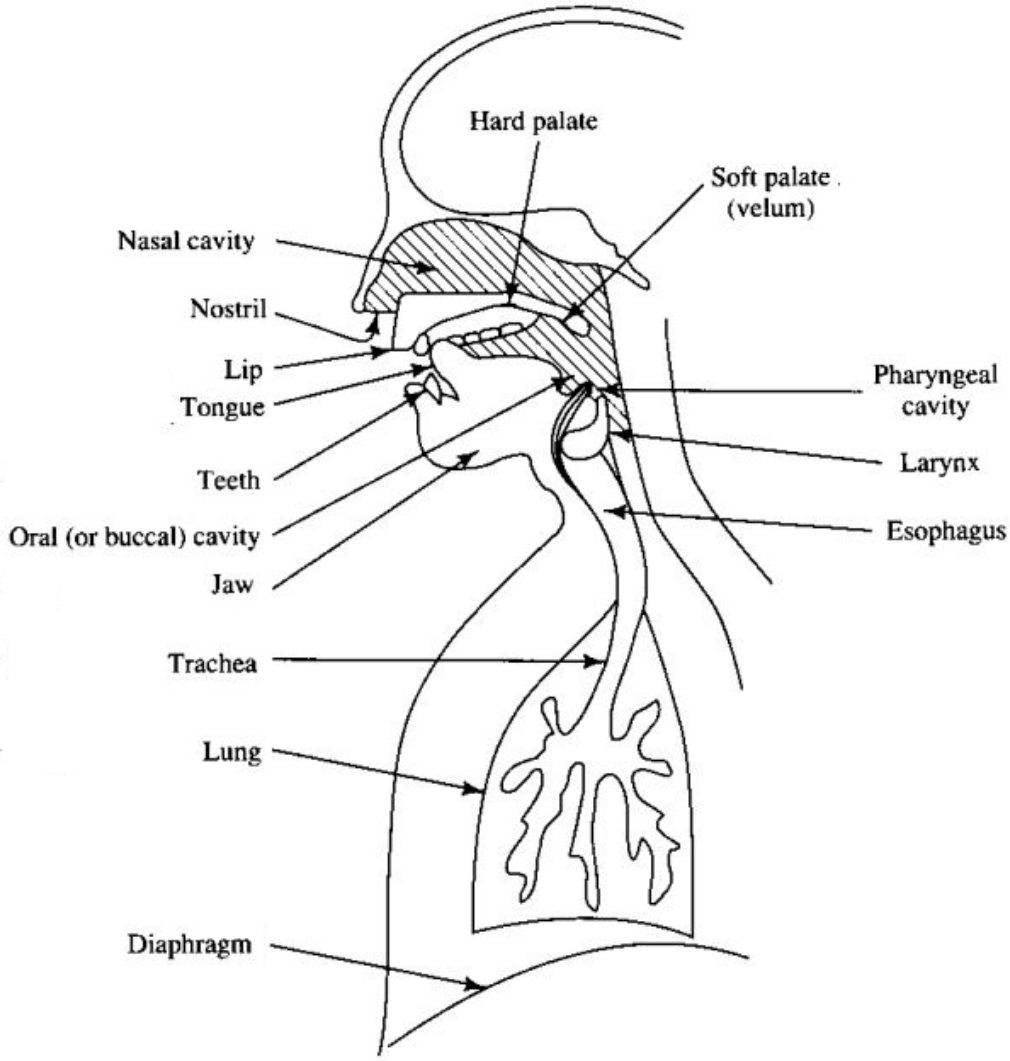
```
python align.py InAudio.wav Transcriptions.txt OutName.TextGrid
```

More information of the toolkit can be found on the following website: `http://www.ling.upenn.edu/phonetics/old_website_2015/p2fa/index.html`.

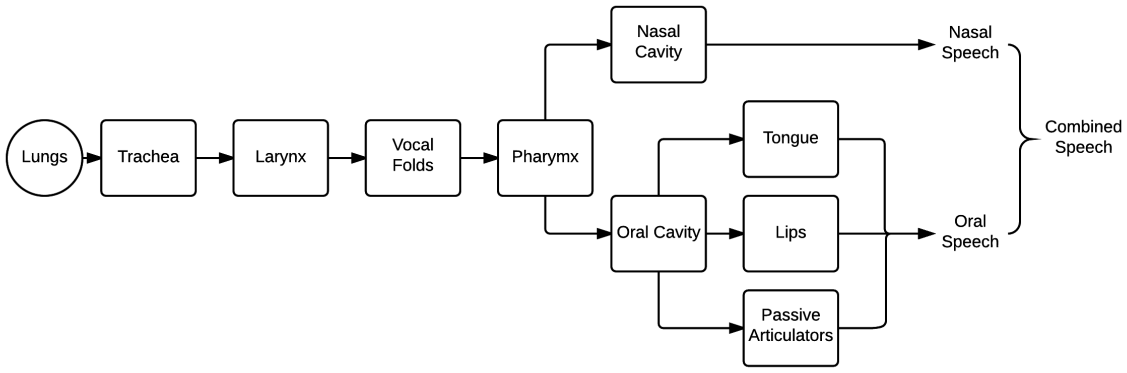
In this work, the P2FA toolkit was used to help find the vowel regions of the utterances. Human verification is done afterwards to ensure correct alignment for each audio signal. Less than 5% of the data requires correction.

2.3 The APP system

The time domain aperiodicity, periodicity, and pitch (APP) detector was built by Deshmukh *et al.* [9] to estimate (1) the proportion of periodic and aperiodic



(a) cross-section view of human vocal system



(b) Speech production mechanism

Figure 2.1: (a) shows a cross-section view of human vocal system and (b) depicts how speech is produced

energy in a speech signal and (2) the pitch period of the periodic component. The APP system uses a time domain method and estimates the pitch based on the distribution of the local minima in the short-time average magnitude difference function (AMDF) of the speech signal. The block diagram of the system is shown in Figure 2.2.

The AMDF [25] is defined as

$$\gamma_n(k) = \sum_{m=-\infty}^{\infty} |x(n+m)w(m) - x(n+m-k)w(m-k)| \tag{2.1}$$

where $x(n)$ is the input signal, $w(m)$ in this APP system is a 20-ms rectangular window and k is the lag value, which varies from 1 to 320 due to the 16kHz sampling rate. This function looks roughly like an inverted autocorrelation function. For truly periodic sounds, the AMDF attains local minima(referred to as dips hereafter) at lags equal to the pitch period and its integer multiples.

If the signal $x(n)$ is truly periodic, it can be seen from equation (2.1) that when k equals a pitch period or multiple of a pitch period, $\gamma_n(k)$ will be brought to zero. That is, there will be spikes sitting at one pitch period, two pitch periods, three pitch periods, etc. in the dip profile. However, since speech is a time-varying process, it is only quasiperiodic in that the pitch period can change somewhat between cycles, and the amplitude of the waveform from one cycle to the next may also change.

The main results from the APP system used in this work is the dip profile. The dip profile stores information about strength and location of dips for every single frame among 60 channels (ERBFilter Bank)[31]. The dip locations and their

strengths, found by computing the convex hull of the AMDF, are applied to determine periodicity and aperiodicity. For a typical periodic frame, the dip profile would show evenly spaced dip clusters, as can be seen in Figure 2.3(a). On the other hand, for an aperiodic frame, dips spread everywhere and their strengths are low, as shown in Figure 2.3(b). Furthermore, the distribution and strengths of the dips can be used to compute the proportion of periodicity and aperiodicity as well as the proportion of periodic and aperiodic energies.

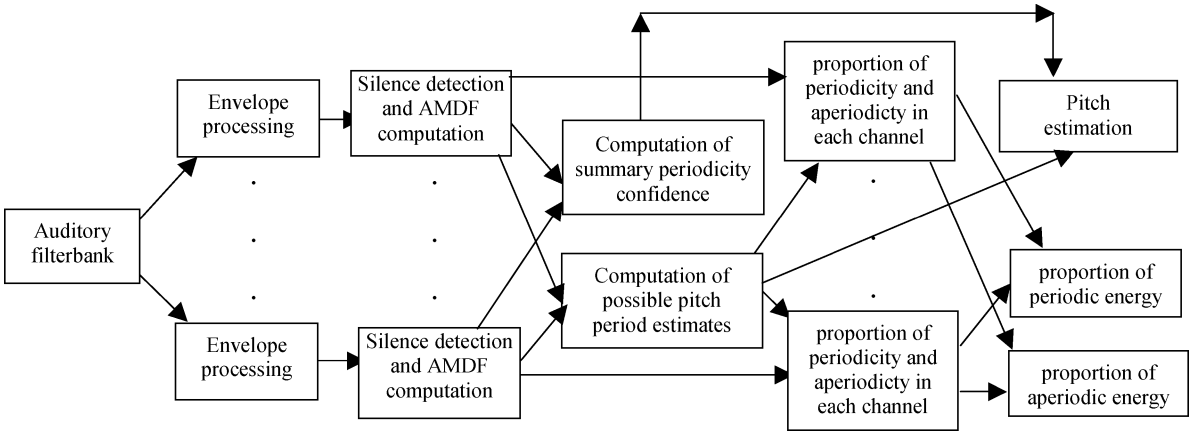
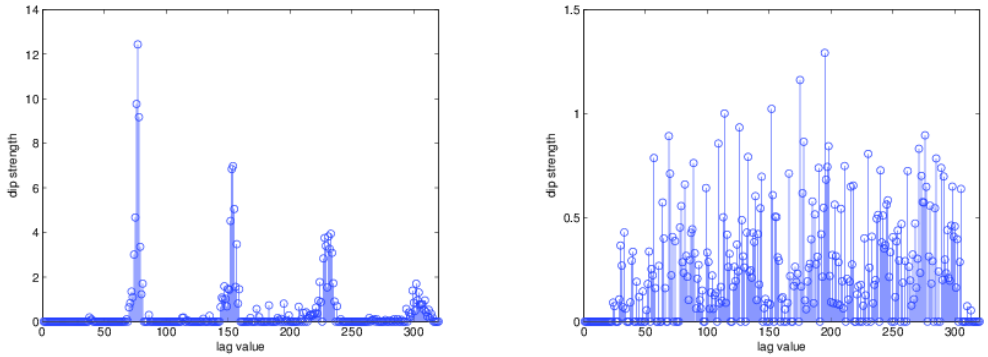


Figure 2.2: Block diagram of APP system. Adapted from "Use of Temporal Information: Detection of Periodicity, Aperiodicity, and Pitch in Speech." by O. Deshmukh, *IEEE Transactions on Speech and Audio Processing*, 13.5 (2005): 776-786.

2.4 Voice Quality

Voice quality is defined by Trask [32] as the characteristic auditory coloring of an individual's voice, derived from a variety of laryngeal and supralaryngeal features and running continuously through the individual's speech. The natural and distinctive tone of speech sounds produced by a particular person yields a particular voice. In this thesis, we are exceptionally interested in jitter, shimmer,



(a) a periodic frame (b) an aperiodic frame

Figure 2.3: Comparison of AMDF dips in periodic and aperiodic frames. (a) shows evenly spaced dip clusters, while (b) shows that dips spread everywhere. Note the huge difference between their strengths.

and breathiness.

2.4.1 Jitter and Shimmer

Jitter and shimmer are measures of the cycle-to-cycle variations of fundamental frequency and amplitude, respectively. By definition,

$$Jitter = \frac{|T_i - T_{i-1}|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \tag{2.2}$$

$$Shimmer = \frac{|A_i - A_{i-1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \tag{2.3}$$

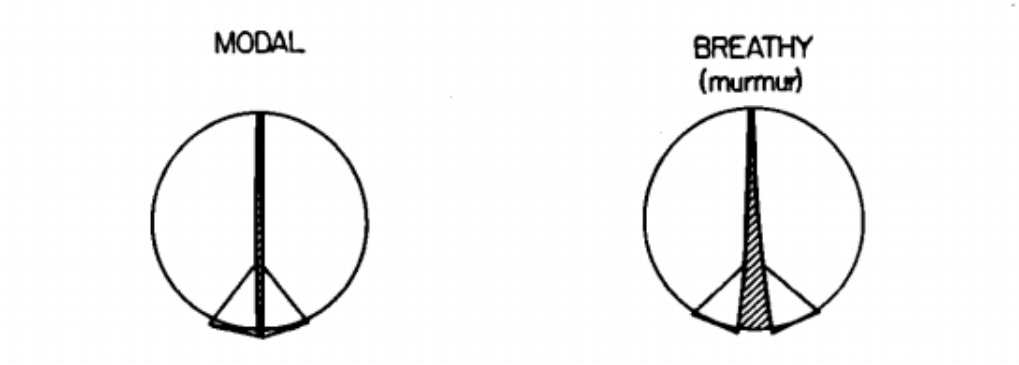
where T_i is the pitch period of the i-th window, A_i is the peak amplitude of the i-th window, and N is the total number of voiced frames. These two parameters can be analyzed under a steady voice producing a vowel continuously.

Vocal jitter is affected mainly by the lack of control of vibration of the cords; it

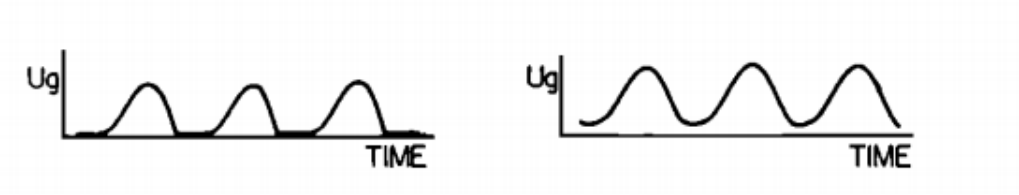
increases in voice disorder and is responsible for hoarse, harsh or rough voice quality. The voices of patients with pathologies often have a higher percentage of jitter. The shimmer changes with the reduction of glottal resistance and mass lesions on the vocal cords and is correlated with the presence of noise emission and breathiness, and it sounds crackly and buzzy. Multi-Dimensional Voice Program (MDVP), a software tool for quantitative acoustic assessment of voice quality, indicates a threshold of pathology of 1.04% for jitter and 3.81% for shimmer. Note that normal voice usually has some amount of jitter and shimmer as they make the voice sound more natural. If they don't vary from cycle to cycle, the voice sounds robotic.

2.4.2 Breathiness

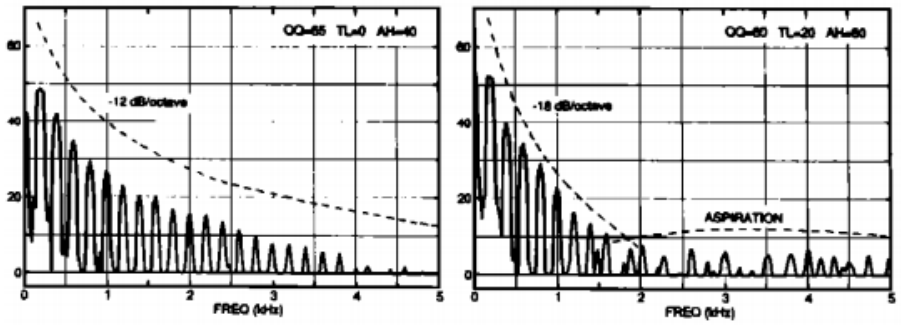
Breathy voice (also called murmured voice) is a phonation in which the vocal cords vibrate, as they do in normal voicing, but don't close along their full length (see Figure 2.4(a)). Muscular tension is low, with minimal adductive tension, weak medial compression and medium longitudinal tension of the vocal folds. Vocal fold vibration is inefficient and, because of the incomplete closure of the glottis, a constant glottal leakage occurs which causes the production of audible friction noise (see Figure 2.4(b)). Aspiration noise can be found in higher frequency region in the spectrogram (see Figure 2.4(c)). Figure 2.5 shows different amount of breathiness on the basis of dip profiles. The dips are higher in the off-peak region when the speech is breathy, as shown in Figure 2.5(b), and vise versa.



(a) Glottal configurations

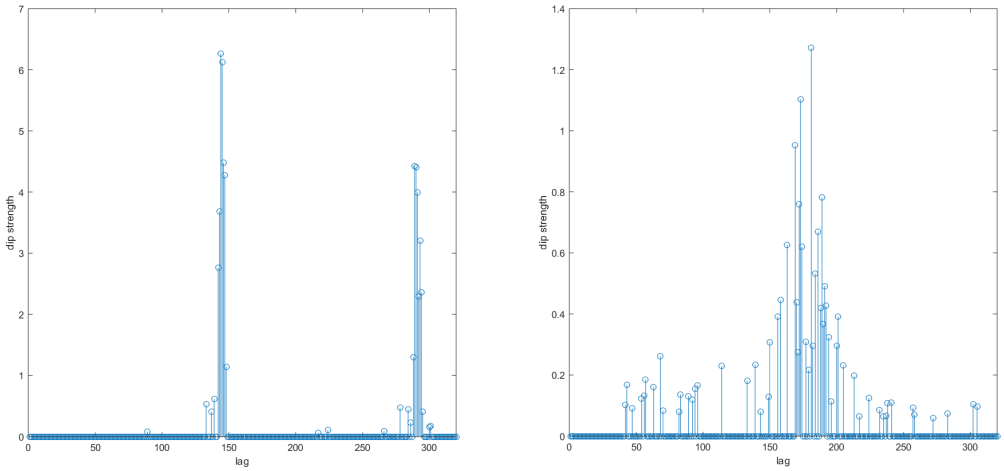


(b) Glottal volume velocity waveforms



(c) Spectrum

Figure 2.4: Glottal configurations and spectrums for modal and breathy voices. Left: modal voice. Right: breathy voice.
(a) shows that vocal folds do not close along their full length when breathy (b) shows there exists a DC component in glottal volume velocity waveform resulting from the incomplete closure of vocal folds in breathy voice (c) shows there is aspiration noise in the higher frequency region when breathy.
This figure is adapted from "Analysis, synthesis, and perception of voice quality variations among female and male talkers." by D. H. Klatt and L. C. Klatt, *the Journal of the Acoustical Society of America*, 87.2 (1990): 820-857.



(a) modal voice (b) breathy voice

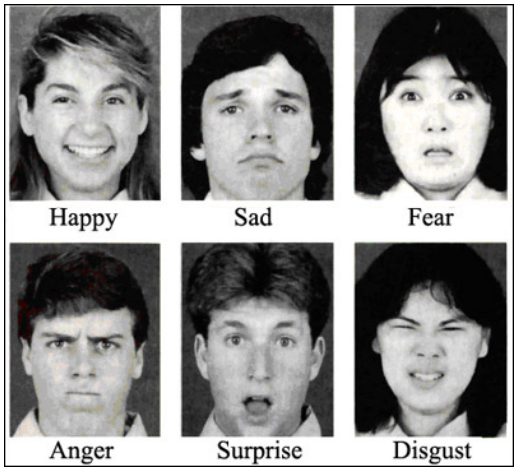
Figure 2.5: This figure shows that the dips are higher in the off-peak region when the voice is breathy. (a) The peaks are "cleaner" (b) Lots of spurious spikes

2.5 Emotions

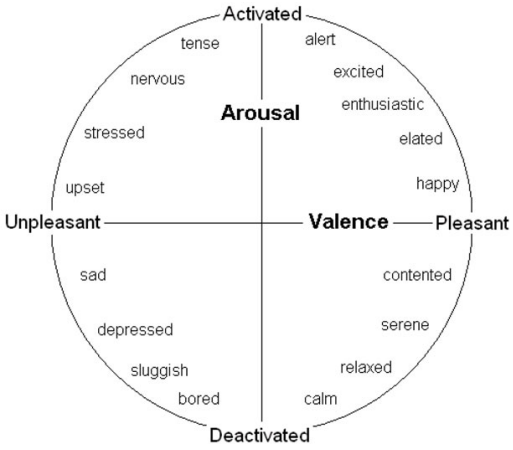
The classification of emotions has been researched from two fundamental viewpoints: (1) emotions are discrete and fundamentally different constructs; or (2) emotions can be characterized on a dimensional basis in groupings. Paul Ekman(1971)[10] devised his list of basic emotions after doing research on many different cultures. He found a high agreement across members of diverse Western and Eastern literate cultures on selecting emotional labels that fit facial expressions. The six basic emotions are: anger, disgust, fear, happiness, sadness, and surprise. James Russell(1980)[26] introduced the circumplex model and proposed that emotions are distributed in a two-dimensional circular space, with arousal and valence being the two axes of the plane. Arousal is the physiological and psychological state of being reactive to stimuli. It results in an observable change in the physical state of the body which causes

you to become alert and a ready to move and respond. Valence means the intrinsic attractiveness (positive valence) or aversiveness (negative valence) of an event, object, or situation. In this model, emotional states can be represented at any level of valence and arousal. Robert Plutchik's(1984)[24] three-dimensional circumplex model, wheel of emotions, describes the relations among emotion concepts, which are analogous to the colors on a color wheel. The cones vertical dimension represents intensity, and the circle indicates degrees of similarity among the emotions. It demonstrates how different emotions can be blend into one another and create new emotions. Plutchik first suggested 8 primary bipolar emotions: joy versus sadness; anger versus fear; trust versus disgust; and surprise versus anticipation.

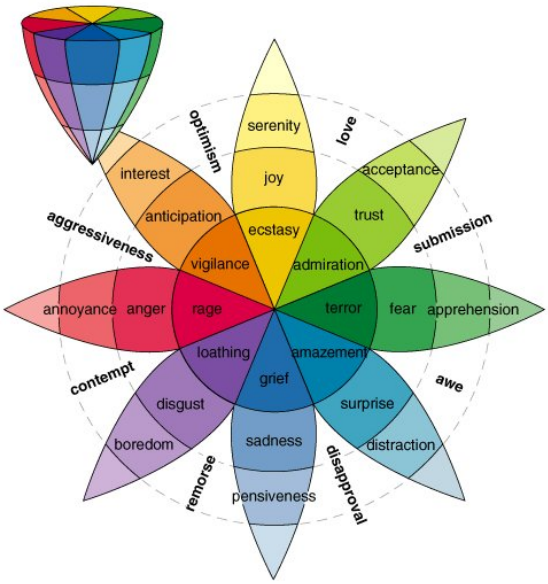
Currently, discrete emotion classification is the state-of-the-art in this area. That being said, we only select finite diverse emotions for classification task. In our case, we strive to classify among angry, happy, neutral, and sad emotions. However, the other models are still useful. For instance, researchers have claimed that global features are efficient only in distinguishing between high-arousal emotions versus low-arousal ones, and this explains why it is harder to tell angry from happy as opposed to angry from sad.



(a) Ekman's six basic emotions ©Paul Ekman



(b) Russell's two dimensional circumplex model



(c) Plutchik's three dimensional circumplex model

Figure 2.6: This figure illustrates three different models for emotion classification as described in section 2.5

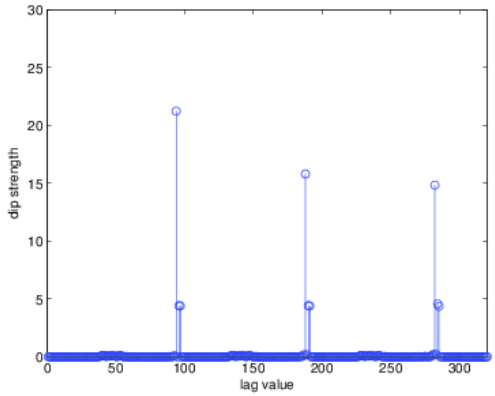
Chapter 3: Study of Emotion Cues from Speech Parameters

3.1 Jitter and Shimmer

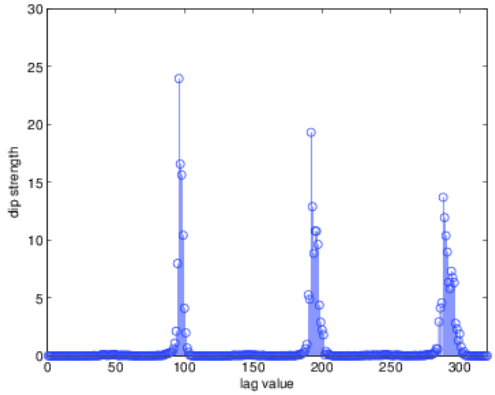
Fundamental frequency F0 is often used in voice assessment. And measurements of F0 disturbance, jitter and shimmer, has proven to be useful in describing vocal characteristics. In this thesis, in place of the jitter/shimmer formulas, the APP detector was used to get the dip profiles to quantify jitter and shimmer. Because equation (2.2) and equation (2.3) only compare the variability within two consecutive periods, while the AMDF equation compares multiple periods as a whole which characterizes the trend more accurately. Based on the meaning of the AMDF equation, jitter and shimmer can be interpreted as the spread and the height of the first dip cluster in the dip profiles during vowel region, respectively. The wider the dip cluster, the more the jitter; the higher the cluster height, the less the shimmer. Figure 3.1 and Figure 3.2 support the statement.

In this work, we look at 14 dips that precede/follow the first peak and calculate jitter as the spread of those dips that exceed 20% of the peak value. Figure 3.3 depicts the idea.

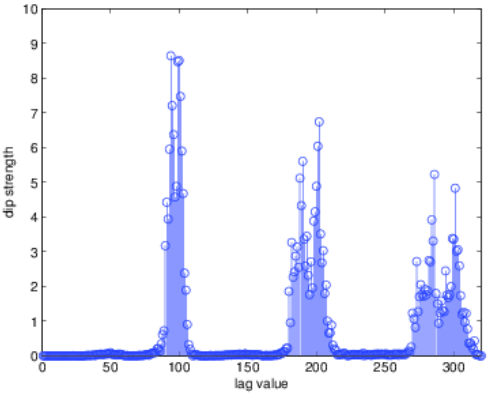
For shimmer, we average the heights of the first cluster in the dip profiles across each frame. To automate the process of finding the peak value in the first dip



(a) /AA/ without jitter

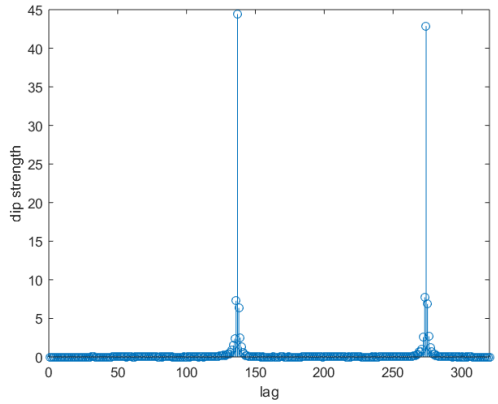


(b) /AA/ with little jitter

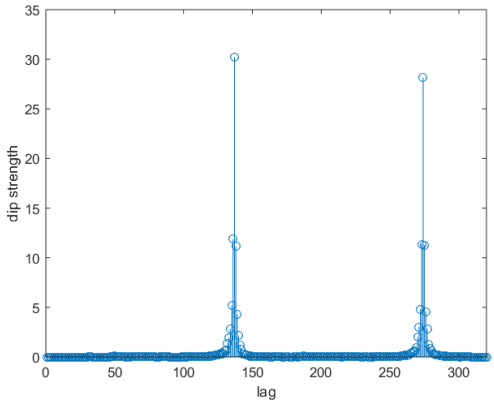


(c) /AA/ with much jitter

Figure 3.1: This figure shows that the amount of jitter is positively correlated with the width of the first dip cluster. One period of /AA/ was first extracted from a real speech. (a) was generated by directly concatenating the signal itself multiple times. Jitter was introduced in (b) and (c) by resampling the signal (to modify pitch period) and concatenation.



(a) /AA/ with little shimmer



(b) /AA/ with much shimmer

Figure 3.2: This figure confirms that the amount of shimmer is negatively correlated with the height of the first dip cluster. One period of /AA/ was first extracted from a real speech. Shimmer was generated in (a) and (b) by scaling the amplitude of that period of /AA/ with random factors and were put together.

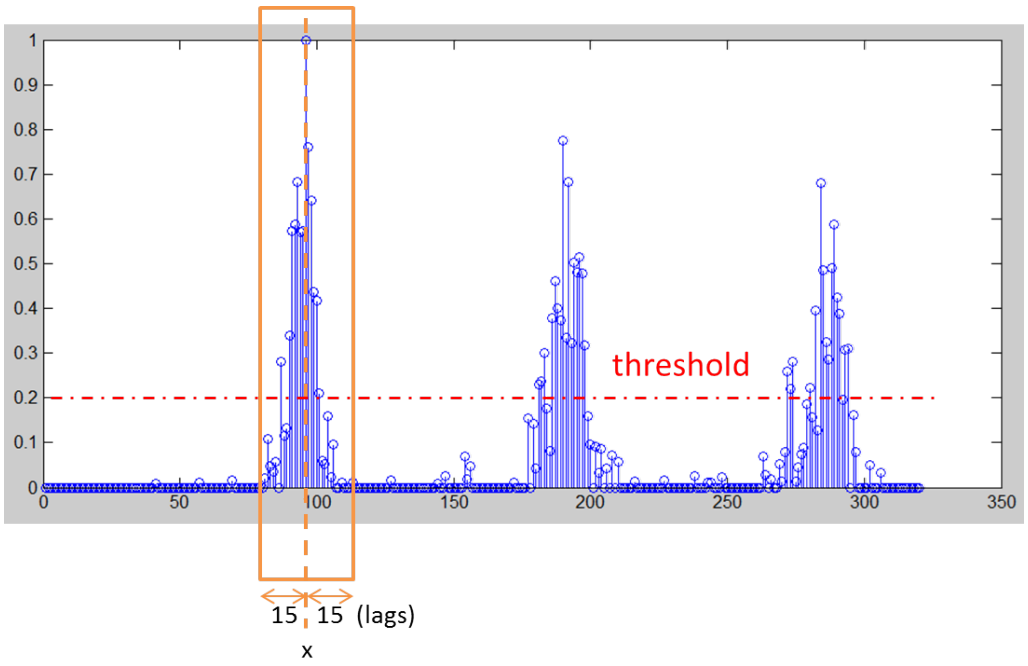


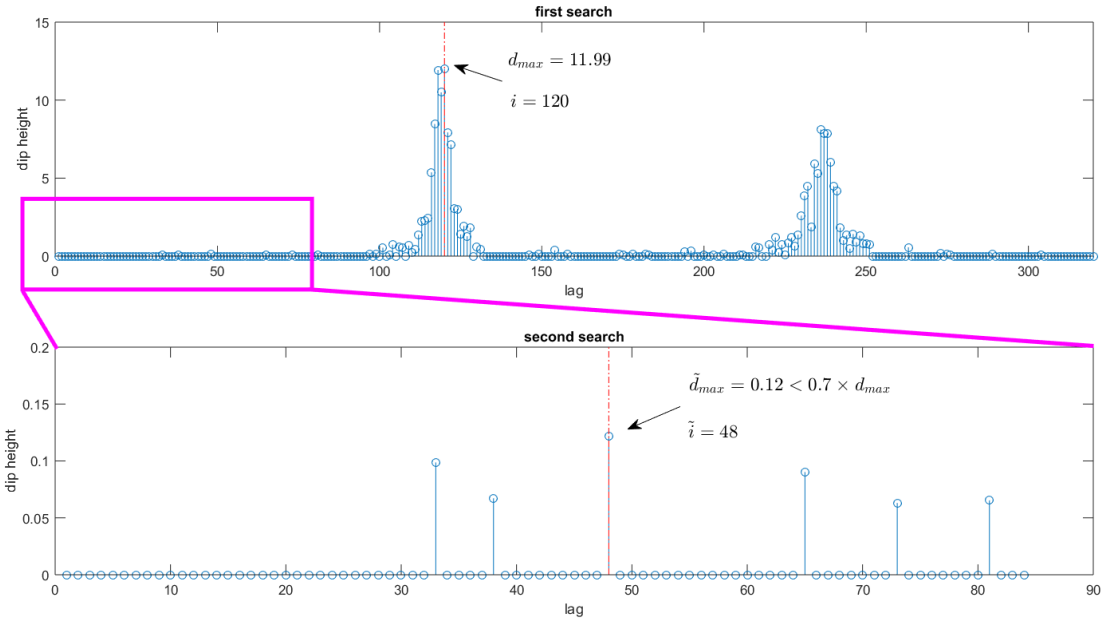
Figure 3.3: This figure demonstrate how we quantify jitter from a dip profile. A window(orange rectangle) of width 29 is centered at the first peak. Standard deviation of distribution of those dips that exceed 20% peak value is said to be the jitter value for this particular speech frame. In this example, the peak occurs at index 96, and the indices of dips that exceeds the threshold are: 87, 90 to 101. Then the spread is calculated as std(87,90,91,92,93,94,95,96,97,98,99,100,101) = 4.182

cluster, the following steps are taken (see Algorithm 1): (1) Search for the maximum value (d_{max}) and its location (i) in the dip profile. (2) Search for the maximum value ($\widetilde{d}_{max}$) and its location only in the first $\lceil 0.7 \times i \rceil$ indices. (3) Set the peak value to $\widetilde{d}_{max}$ if $\widetilde{d}_{max} \geq 0.7 \times d_{max}$, we conclude this new maximum value to be the peak value for the first dip cluster; this happens, though rarely, when the second cluster has a higher peak value than the first cluster. (4) Otherwise, set peak value to d_{max} when $\widetilde{d}_{max} < 0.7 \times d_{max}$. The threshold of 0.7 was chosen empirically. Two example dip profiles are shown in Figure 3.4.

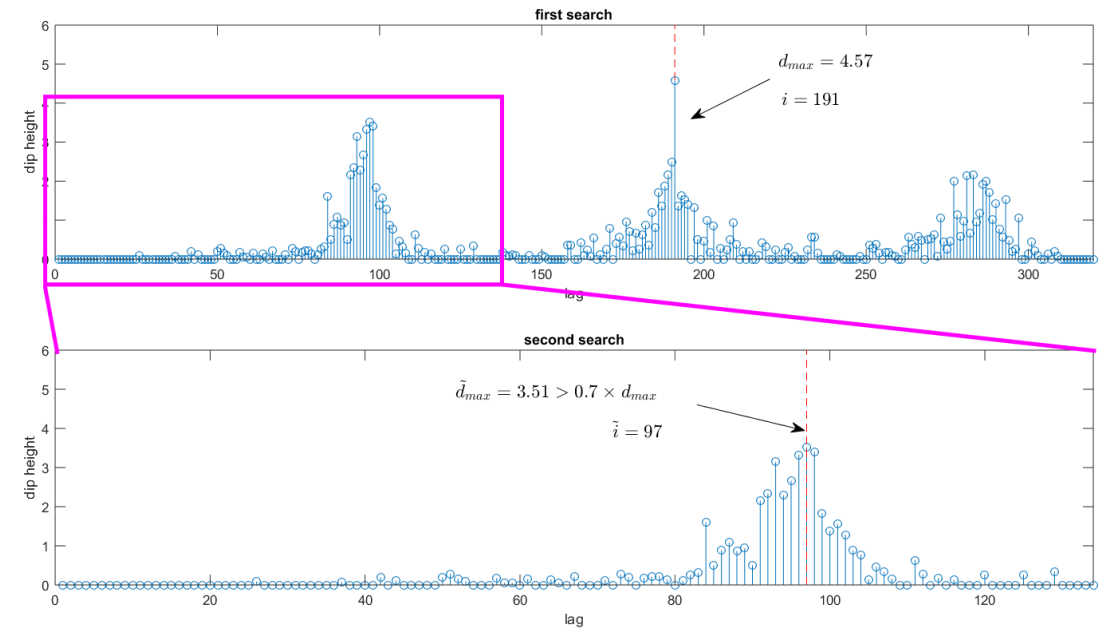
Algorithm 1 Algorithm for automatically finding peak value in the first dip cluster

```
function getSHIMMER(dipprofile)
    n = length(dipprofile)
    [d_max, i] = max(dipprofile(1:n))   ▷ find the largest element and its index in dipprofile
    [d̃_max, ĩ] = max(dipprofile(1:⌈0.7 × i⌉))   ▷ find the largest element and its index within the first ⌈0.7 × i⌉ samples in dipprofile
    if d̃_max ≥ 0.7 × d_max then
        return d̃_max
    else
        return d_max
    end if
end function
```

We explored jitter and shimmer produced by different subjects with various emotions and found out that speech produced with neutral and sad emotions tends to have higher jitter values relative to speech produced with angry and happy emotions. And shimmer is lower for angry, happy and sad speech, relative to neutral speech. Figure 3.5 shows an example boxplot for jitter and Figure 3.6 shows an example boxplot for shimmer, both with four emotions: angry, happy, neutral, and sad. The boxplot is a convenient way of graphically depicting groups of numerical data



(a) Max value occurs in first dip cluster, shimmer = 11.99



(b) Max value occurs in second dip cluster, shimmer = 3.51

Figure 3.4: This figure illustrates how our algorithm works to capture the height of the first dip cluster (a) shows when the max dip occurs in the first cluster while (b) shows when the max dip occurs in the second cluster, the algorithm will update the selection by checking dips preceeding the maximum.

through their quartiles and gives us a sense of how the data distribute. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually.

3.2 Breathiness/Aperiodic Energy During Vowels

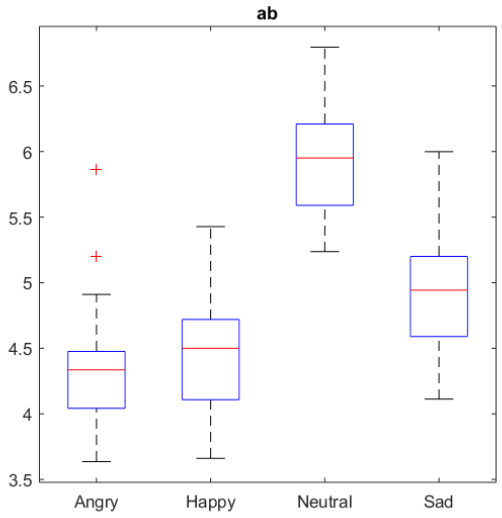
Breathy voice comes from incomplete closure of the vocal folds and appears as aspiration noise in the higher frequency range. (see Figure 2.4(c)) Aperiodic energy during vowels could be used to assess the amount of breathiness in the higher-frequency region. This parameter is believed to contain emotion-related information. In particular, we found that when people are sad, their voices tend to be breathier. Scherer(1986)[27] suggests that lax voice (speech sound pronounced with little muscular effort and consequently having relatively imprecise accuracy of articulation and little temporal duration; at the phonatory level essentially the same as breathy voice) is associated with sadness. Laukkanen *et al.*(1996)[17] indicates that sorrow was produced with a more breathy voice quality. Burkhardt and Sendlmeier(2000)[4] describe a synthesis system for the generation of emotional speech. They found that breathy voice is associated with sadness.

To measure breathiness, the APP detector was once again used to get the dip profiles, and dips occurring outside a certain tolerance region of the clusters are summed together. Note that breathiness is aperiodicity in higher frequency channels, so we only sum the dips for the channels above 2500Hz. The greater the

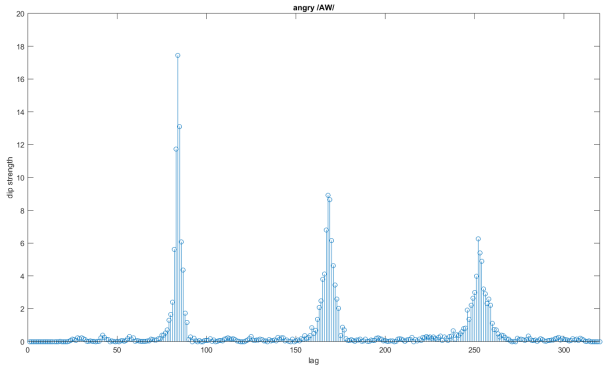
sum, the breathier the voice. Figure 3.7(a) shows speech produced with neutral and sad emotions have more aperiodic energy relative to speech produced with angry and happy emotions. Figure 3.7(b) shows sample dip profiles for angry and sad speech. The dips are relatively higher in the off-peak region in sad speech.

3.3 Speaking Rate

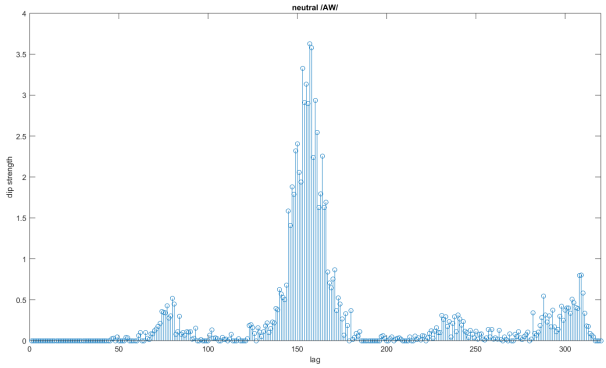
Speaking rate is supposed to reflect the speed at which an individual executes articulatory movements for speech production. Williams and Stevens (1972)[35] found higher syllabic rate in neutral (4.31 syllables per second) and anger (4.15 syllables per second), while lower rate in sorrow (1.91 syllables per second). There were longer vowels and consonants and longer pauses in sorrow, relative to neutral and anger, that were often inserted in a sentence. Braun and Oba (2007) [2] discovered greatest syllable rate in hot anger, and lowest in sadness. In this work, speaking rate was assessed as articulation rate, as number of syllables divided by total duration of speaking time. Approximately 80% of the utterances consist of single sentence, while the rest 20% consist of two sentences, and pause between the two sentences are excluded. Speech data used in this work is scripted and the CMU Pronouncing Dictionary was used to obtain number of syllables(vowels). Utterance durations were measured from the corresponding label files produced by the Penn Phonetics Lab Forced Aligner [38]. Our study shows highest rate in neutral speech and lowest rate produced with sad emotion as illustrated in Figure 3.8.



(a) distribution of jitter under different emotions



(b) dip profile for angry /AW/



(c) dip profile for neutral /AW/

Figure 3.5: (a) shows the boxplot of jitter for speaker **ab** under four different emotions. It can be seen that speech produced with neutral and sad emotions have higher jitter values relative to speech produced with angry and happy emotions. (b) and (c) are sample dip profiles from speaker **ab**. The spread of dips are wider in neutral than in angry.

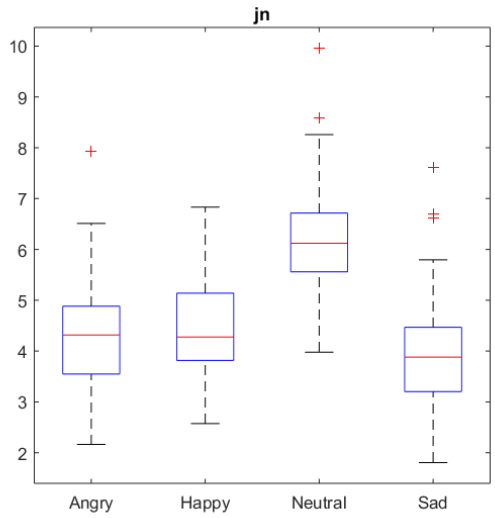
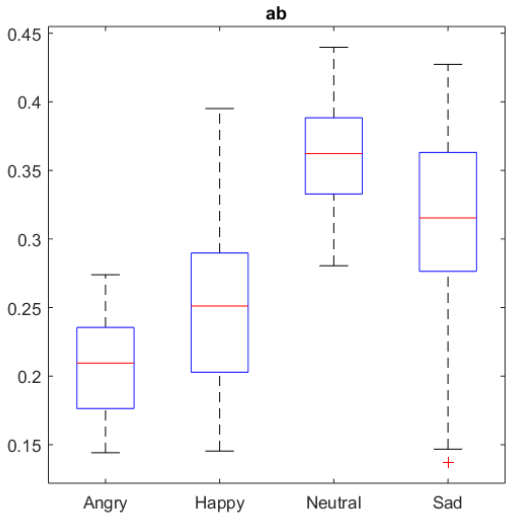
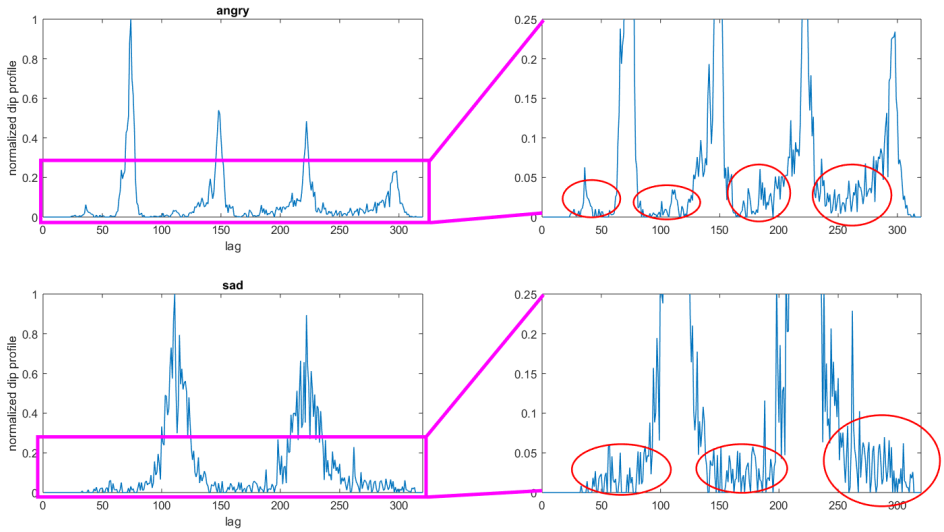


Figure 3.6: This figure shows the boxplot of shimmer for speaker **jn** under four different emotions. It shows shimmer being higher in neutral speech, relative to the other three.



(a) Boxplot of breathiness for speaker **ab** under four different emotions.



(b) dip profiles: angry vs sad

Figure 3.7: (a) indicates speech produced with neutral and sad emotions have more aperiodic energy relative to speech produced with angry and happy emotions. (b) shows the normalized dip profiles from an angry and a sad speech. From the figure, it can be seen that the dips are relatively higher in the off-peak region (circled in red) in sad speech.

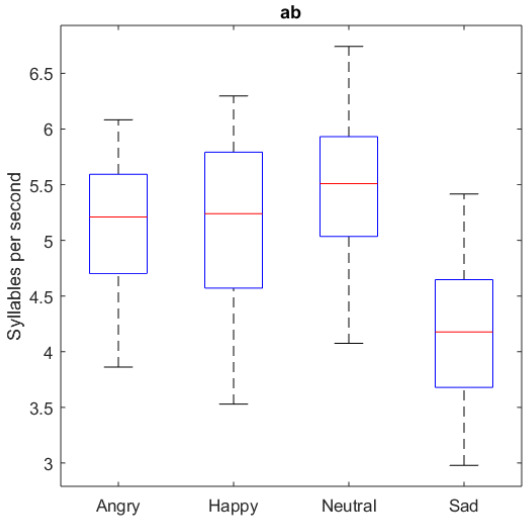


Figure 3.8: This figure shows the boxplot of articulation rate for speaker **ab** under four different emotions. It suggests fastest speaking rate in neutral speech, and slowest rate in sad speech.

Chapter 4: Materials, Methods, and Results

4.1 Speech Corpus

The speech data used in this study is part of the Electromagnetic Articulography (EMA) database. The EMA database is made of acted emotions and was collected in a nonechoic chamber by the Signal Analysis and Interpretation Laboratory (SAIL) at the University of Southern California in year 2005. Three speakers (1 male and 2 females) were asked to repeated 10 or 14 sentences five times each in a random order. The sentences are mostly neutral in emotional content (see Table 4.1). Four different emotions: angry happy, neutral, and sad, were simulated, resulting in a total of 200 or 280 utterances (10 or 14 sentences $\times$ 5 repetitions $\times$ 4 emotions) for each subject. Each utterance was then digitized in 12-bit resolution with 16kHz sampling rate. The EMA data was collected simultaneously that track the positions of three sensors in the midsagittal plane adhered to the tongue tip, the lower maxilla and the lower lip. Given the EMA database consists of the same sentence produced many times by the same subjects where the only difference was the emotion expressed, and the database consists of clean speehch in American English, it allows for a controlled study. More details about the dataset can be found in [18].

Considering the difficulty in getting articulatory information in real life, we focus only on the acoustic data from this database. To determine how well the data represents each emotional state, SAIL conducted human evaluation tests with 4 English speakers. They listened to the recordings and were asked to judge the emotion expressed in each utterance.

The utterances which obtain both high rates (3 or 4) for its target emotion and low rates (0 or 1) for the other emotions were selected for this study. A total number of 507 utterancess were chosen, and 295 of them are perfect emotion utterances (i.e. all the evaluators chose the same target emotion). The distribution of data is presented in Table 4.2.

#	**Sentence**
1.	Your grandmother is on the phone.
2.	Don't compare me to your father.
3.	I hear the echo of voices and the sound of shoes.
4.	That dress looks like it comes from Asia.
5.	They think the company and I will have a long future.
6.	The doctor made the scar. Foam antiseptic didn't help.
7.	That made being deaf tantamount to isolation.
8.	The doctor made the scar foam with antiseptic.
9.	I am talking about the same picture you showed me.
10.	It's hard being very deaf. Tantamount to isolation.

Table 4.1: The list of 10 sentences from the EMA database used in this study

4.2 Experiments

4.2.1 Methodology

The overall procedures of a recognition task are as follows: The speech data are first trimmed (based on a silence detector) so that the silence part in the beginning

subject	**ab**		**jn**		**ls**		
	perfect	others	perfect	others	perfect	others	
Number of Utterances	14	22	21	23	34	14	Angry
	15	18	23	17	20	17	Happy
	20	19	39	10	27	17	Neutral
	32	14	20	23	30	18	Sad
	81	**73**	**103**	**73**	**111**	**66**	**Total**

Table 4.2: This table displays the distribution of our speech data. "perfect" indicates number of utterances which all human evaluators assess the same as their target emotion. "other" means 3 of the evaluators judged the utterance as its target emotion, while the other evaluator chose a different emotion.

and the end of the speech are removed. Acoustic features are then extracted using several methods. Afterwards, a classifier is used and 10-fold cross validation is conducted to evaluate the performance of the recognizer. The detailed procedures are shown in Figure 4.1.

4.2.2 Feature Extraction

The common trend of feature extraction consists of extracting key features from speech samples and creating long vectors. In this work, the open-Source Media Interpretation by the Large feature-space Extraction (openSMILE) toolkit [12] is used for most of the feature extraction process. The toolkit is capable of extracting many helpful acoustic features, called low-level descriptpors (LLD), and their statistics for distinct purposes of speech recognition tasks. With divergent configuration settings, openSMILE will extract different sets of LLD accordingly.

In previous work, researchers investigated the use of various acoustic parameters for emotion recognition. These parameters include pitch, loudness, mel-frequency cepstral coefficients, and speaking rate etc. The configuration used in

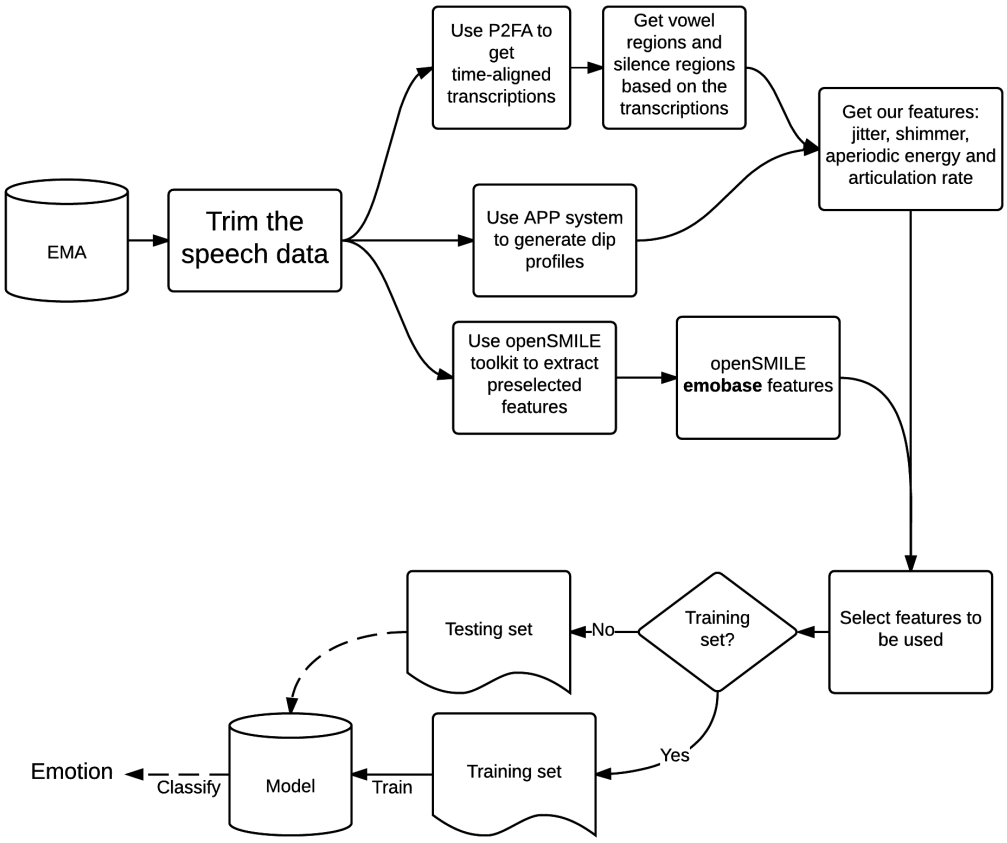


Figure 4.1: The flowchart depicts the big picture of how a recognition task is done.

this study is the openSMILE **'emobase'** set, which is a standard baseline set designed for emotion recognition tasks. A total of 988 acoustic features for emotion recognition can be extracted.

The feature set specified by `emobase.conf` contains the following LLD: Intensity, Loudness, 12 Mel-Frequency Cepstrum Coefficients(MFCCs), Pitch(F0), Probability of voicing, F0 envelope, 8 Line Spectral Frequencies(LSFs), Zero-Crossing Rate. Delta regression coefficients are computed from these LLD, and the following functionals are applied to the LLD and the delta coefficients: Max./Min. value and respective relative position within input, range, arithmetic mean, 2 linear regression coefficients and linear and quadratic error, standard deviation, skewness, kurtosis, quartile 1-3, and 3 inter-quartile ranges.

Apart from the 988 features obtained from the openSMILE toolkit, another 8 features are extracted using our methods descirbed in Chapter 3. They are maximum, minumum and mean of jitter and shimmer, aperiodic energy measure during vowels, and average articulation rate.

4.2.3 Classifier

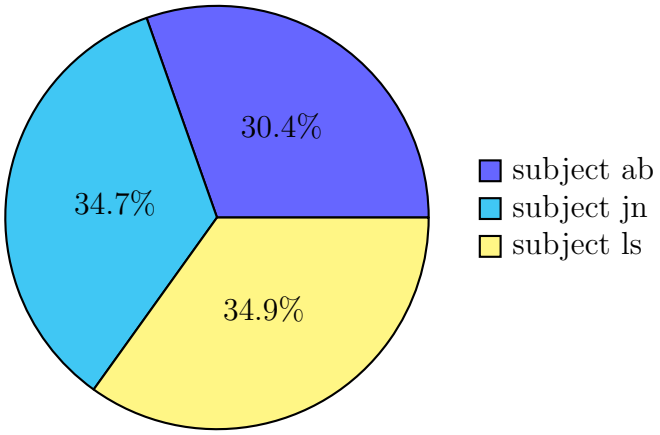
The 5-Nearest Neighbor classifiers with Euclidean distance is used for training and testing. Essentially, every single testing data is classified by a majority vote of its closest 5 neighbors in the training set. The number 5 is chosen empirically. Since the computational complexity for the nearest-neighbor algorithm is high in both space and time, not to mention the openSMILE emobase feature set being considerably

high dimensional (988), we were also motivated to reduce the dimensionality of the feature set at the same time as we explored new features.

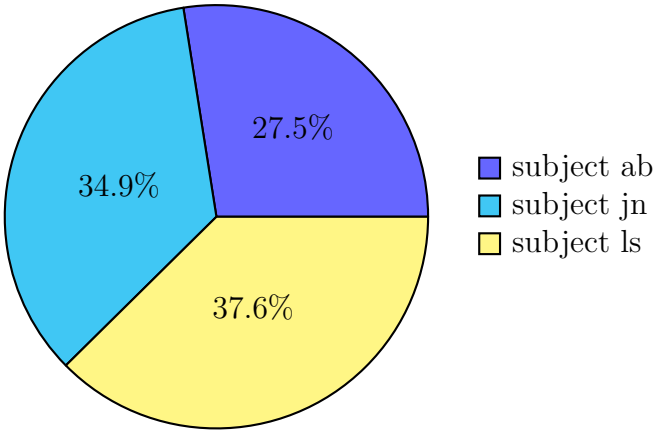
4.3 Methods and Results

The first part of the experiments focused on dimension reduction. Our goal was to figure out a smaller set of features that perform as well as using the whole feature set. The feature set is separated into groups systematically based on their types, and several recognition trials were performed using multiple groups of features. To estimate how the results will generalize to an independent data set, we performed 10-fold cross validation. Basically, the data set is broken into 10 groups with each group approximately the same size. For each epoch, one group of data serve as testing set, and the rest as the training set. The recognition rate is then calculated and averaged over 10 iterations.

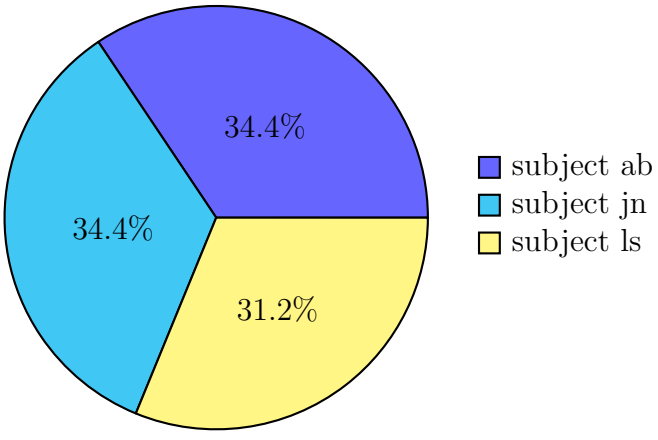
The other part of the experiments is to include our own features—jitter, shimmer, aperiodic energy during vowels, and articulation rate—into the reduced set and see how they perform. In addition to 10-fold cross validation, we also ran the recognition task using the "perfect" utterances as training set, and the rest as testing set. Intuitively, those "perfect" utterances can be treated as the paragon of emotional speech given that all human evaluators rate it the same way. Moreover, the data distribute evenly in both training and testing set with the setting (see Figure 4.2). Thereby it is reasonable to run the experiments using this setting and compare the results.



(a) all 507 utterances



(b) 295 perfect utterances



(c) 212 other utterances

Figure 4.2: The pie charts show the distribution of utterances from different subjects in three groups. As can be seen from the charts, the dataset is pretty much balanced regardless of grouping. This ensures the classification results won't be biased.

In most cases, using only MFCC related features to perform the recognition tasks does not degrade the accuracy; instead the subset outperforms the whole **emobase** feature set. Furthermore, the third row of Table 4.3 shows that with additional F0 related features, the accuracy using the MFCCs can be slightly improved. Thus, We claim that using only MFCC and pitch related features should be sufficient. Aside from the improvement in recognition rate, the computation compelexity is greatly reduced by 72% owing to dimension reduction. What's more, the recognition rate can be further improved by adding our own features. Figure. 4.3 shows an example of recognition rate, assessed by 10-fold cross validation, using various sets of features and their dimensionalities. Using MFCC and pitch related features along with our features attains the highest recognition rate at 87.97% in this example. Huamn judges attain a recognition rate at (295×100%+212×75%)/507=89.55%.

The results of the experiment about using the "perfect" utterances as training set and the rest (human judges achieve 75% recognition rate.) as testing set can be found in Table 4.3. Using our set of features along with the MFCCs and F0 related features come to a recongition rate of 81.13%, which is 3.3% higher than using the whole **'emobase'** feature set, and 6.13% higher than huamn judges. Confusion matrices for **'emobase'** features versus our features are provided in Table 4.4. It is shown that most improvement comes from correctly classified angry speeches. We improve the recognition rate of angry emotion by 10% in paricular.

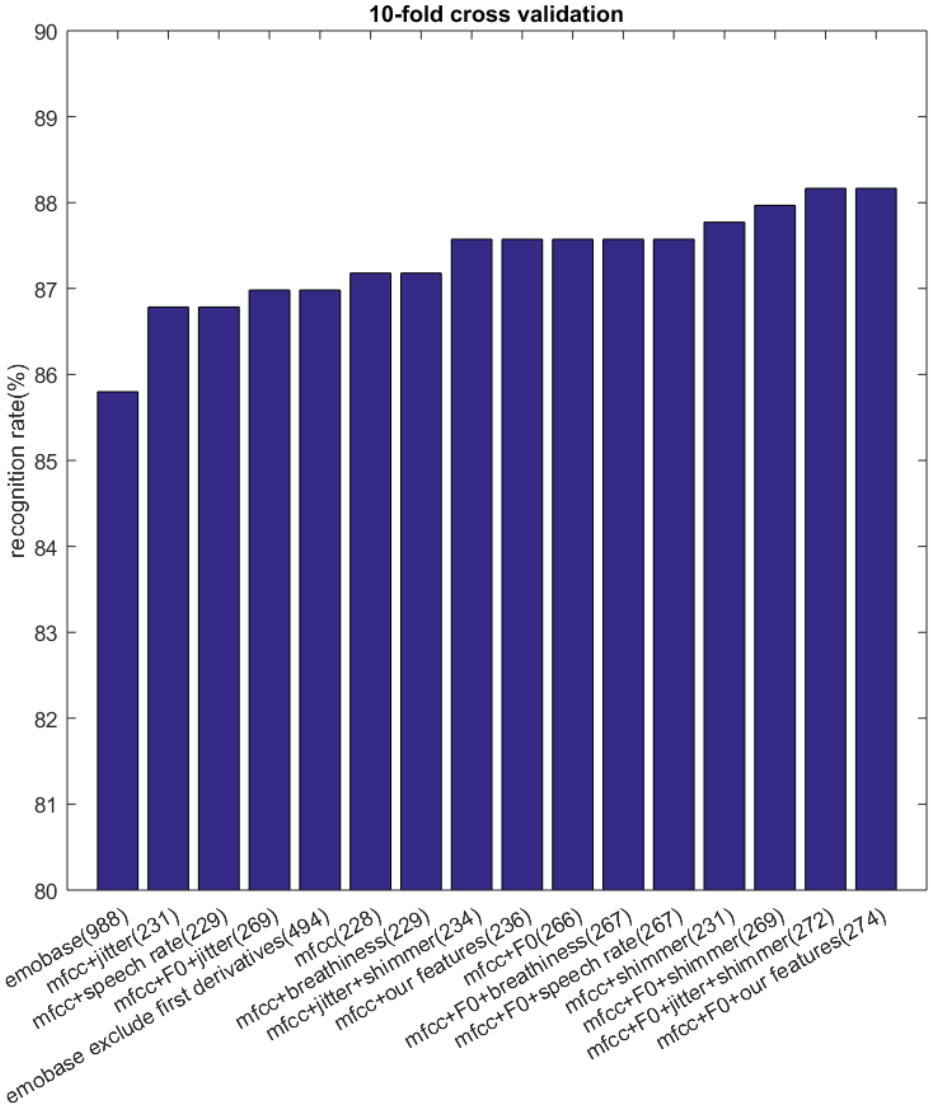


Figure 4.3: This figures shows an example recognition rate assessed using 10-fold cross validation. It can be seen that using only MFCC and pitch related features as well as our features gives the highest recognition rate at 88.17%. In comparison, emobase features achieve 85.8% correctness, MFCC related features reach 87.18%, and MFCC plus pitch related features attain 87.57%. In addition to improvement of recognition rate, the amount of features used is greatly reduced from 988 to less than 280. Note that numbers in the parentheses indicate the dimensionality of that feature set.

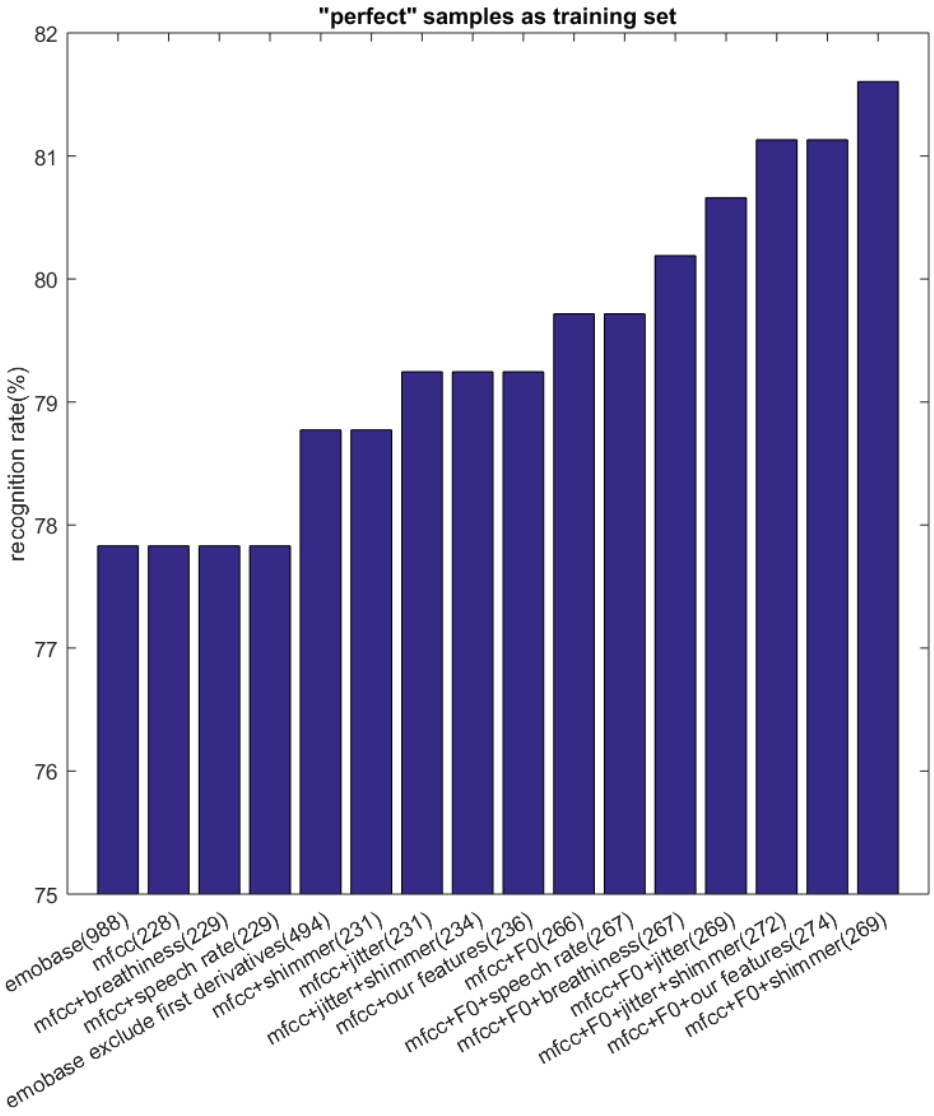


Figure 4.4: This figures displays recognition rates with perfect utterances as training set using various features.It shows that using only selected features can achieve better recognition rate (above 81%) in comparison to using the whole emobase feature set.(77.83%)

	accuracy(%)
Human Judges	75.00
emobase features (988)	77.83
MFCC + F0 (266)	79.72
MFCC + our features(236)	79.25
MFCC + F0 + our features (274)	81.13
MFCC + F0 + shimmer (269)	81.60

Table 4.3: This table displays the recognition accuracy using the "perfect" utterances as training set and the rest as testing set. Our reduced feature set outperforms the openSMLIE emobase feature set by more than 3%.

		predicted				
		Angry	Happy	Neutral	Sad	Accuracy per emotion
actual	Angry	36	10	10	3	61.02%
	Happy	6	42	1	3	80.77%
	Neutral	2	1	43	0	93.48%
	Sad	1	0	10	44	80.00%
					Overall	77.83%

(a) **'emobase'** features

		predicted				
		Angry	Happy	Neutral	Sad	Accuracy per emotion
actual	Angry	42	6	9	2	71.19%
	Happy	6	42	0	4	80.77%
	Neutral	1	0	44	1	95.65%
	Sad	0	0	11	44	80.00%
					Overall	81.13%

(b) MFCC + F0 + our features

Table 4.4: The table shows the confusion matrix when using perfect utterances as training set. From the statistics we can see that the recognition rate for angry speech is improved by 10 %.

Chapter 5: Conclusions

5.1 Summury

The understanding of how speech is affected by emotion is crucial towards improving the performance of an emotion recognition system. Thanks to advances in technology, computers get faster processors and lower prices. People nowadays can easily neglect the importance of computation complexity and perform recognition tasks in a data-driven approach. In spite of that, this thesis deeply analyzes a few features (jitter, shimmer, aperiodic energy, and articulation rate) and finds them to be involved with emotional speech. We show by the EMA dataset that using our features together with MFCC and pitch related features lead to a better performance in comparison to the openSMILE **'emobase'** feature set. It not only improves the overall accuracy by 3.3% but also greatly reduces the feature space by 72%.

5.2 Future Work

While this thesis provides a basic framework for emotion recognition from speech, more work is needed in several areas.

- **Collect more emotional speech data:** Lack of data is always a big problem

in the community. Most of the emotional speech datasets are private or in different languages. We are in need of a standard database so that people can easily compare their results.

- **Investigate new features:** Needless to say, there is still a gap between the performance of speech emotion perception by human and machine. Based on the results we get, our features contribute the most in differentiating angry speech. There could be other features which recognize happy or sad emotions better. The right features may be hard to find and need more exploration.
- **Reduce feature space even more:** In this work, features are selected on a all-or-nothing basis. Take MFCC for example, we either keep all the MFCC related features or none of them. It is possible to reduce the feature space even more if additional investigation could be done on a functional basis. We are still not clear whether the statistics applied are all beneficial or not.
- **Automate and simplify our feature extraction process:** In this work, all the features are extracted separately using different modules, and were concatenated afterwards. The tasks could possibly be integrated into one single package and all the features could be extracted and concatenated at once seamlessly.

Bibliography

[1] Rainer Banse and Klaus R Scherer. Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70(3):614, 1996.

[2] Angelika Braun and Reiko Oba. Speaking tempo in emotional speech–a cross-cultural study using dubbed speech. In *Proceedings 16th International Conference on Phonetic Sciences, Saarbrücken, Germany*, pages 77–82. Citeseer, 2007.

[3] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss. A database of german emotional speech. In *Interspeech*, volume 5, pages 1517–1520, 2005.

[4] Felix Burkhardt and Walter F Sendlmeier. Verification of acoustical correlates of emotional speech using formant-synthesis. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[5] Carlos Busso, Murtaza Bulut, and Shrikanth Narayanan. Toward effective automatic recognition systems of emotion in speech. *Social emotions in nature and artifact: emotions in human and human-computer interaction, J. Gratch and S. Marsella, Eds*, pages 110–127, 2012.

[6] Lijiang Chen, Xia Mao, Yuli Xue, and Lee Lung Cheng. Speech emotion recognition: Features and classification models. *Digital signal processing*, 22(6):1154–1160, 2012.

[7] Nivja H De Jong and Ton Wempe. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods*, 41(2):385–390, 2009.

[8] Frank Dellaert, Thomas Polzin, and Alex Waibel. Recognizing emotion in speech. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1970–1973. IEEE, 1996.

[9] Om Deshmukh, Carol Y Espy-Wilson, Ariel Salomon, and Jawahar Singh. Use of temporal information: Detection of periodicity, aperiodicity, and pitch in speech. *IEEE Transactions on Speech and Audio Processing*, 13(5):776–786, 2005.

[10] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

[11] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.

[12] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the international conference on Multimedia*, pages 1459–1462. ACM, 2010.

[13] Christer Gobl, Ailbhe Nı, et al. The role of voice quality in communicating emotion, mood and attitude. *Speech communication*, 40(1):189–212, 2003.

[14] Dennis H Klatt and Laura C Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *the Journal of the Acoustical Society of America*, 87(2):820–857, 1990.

[15] A Kolakowska, Agnieszka Landowska, Mariusz Szwoch, Wioleta Szwoch, and Michał R Wróbel. Emotion recognition and its application in software engineering. In *2013 The 6th International Conference on Human System Interaction (HSI)*, pages 532–539. IEEE, 2013.

[16] Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao, and Te-Won Lee. Emotion recognition by speech signals. In *In Proceedings of International Conference EUROSPEECH*. Citeseer, 2003.

[17] Anne-Maria Laukkanen, Erkki Vilkman, Paavo Alku, and Hanna Oksanen. Physical variations related to stress and emotional state: a preliminary study. *Journal of Phonetics*, 24(3):313–335, 1996.

[18] Sungbok Lee, Serdar Yildirim, Abe Kazemzadeh, and Shrikanth Narayanan. An articulatory study of emotional speech production. In *INTERSPEECH*, pages 497–500, 2005.

[19] Xi Li, Jidong Tao, Matthew Thomas Johnson, Joseph Soltis, Anne Savage, Kirsten M Leong, and John D Newman. Stress and emotion classification using jitter and shimmer features. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–1081. IEEE, 2007.

[20] Mark Liberman, Kelly Davis, M Grossman, N Martey, and J Bell. Emotional prosody speech and transcripts. *Linguistic Data Consortium, Philadelphia*, 2002.

[21] Iain R Murray and John L Arnott. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, 93(2):1097–1108, 1993.

[22] Joy Nicholson, Kazuhiko Takahashi, and Ryohei Nakatsu. Emotion recognition in speech using neural networks. *Neural computing & applications*, 9(4):290–296, 2000.

[23] Paolo Petta, Catherine Pelachaud, and Roddy Cowie. Emotion-oriented systems. *The Humaine Handbook*, 2011.

[24] Robert Plutchik. Emotions: A general psychoevolutionary theory. *Approaches to emotion*, 1984:197–219, 1984.

[25] M Ross, H Shaffer, Andrew Cohen, Richard Freudberg, and H Manley. Average magnitude difference function pitch extractor. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 22(5):353–362, 1974.

[26] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[27] Klaus R Scherer. Vocal affect expression: a review and a model for future research. *Psychological bulletin*, 99(2):143, 1986.

[28] Björn Schuller, Gerhard Rigoll, and Manfred Lang. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 1, pages I–577. IEEE, 2004.

[29] Björn Schuller, Stefan Steidl, and Anton Batliner. The interspeech 2009 emotion challenge. In *INTERSPEECH*, volume 2009, pages 312–315, 2009.

[30] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian MüLler, and Shrikanth Narayanan. Paralinguistics in speech and language—state-of-the-art and the challenge. *Computer Speech & Language*, 27(1):4–39, 2013.

[31] Malcolm Slaney. Auditory toolbox version 2, 1998.

[32] Robert Lawrence Trask. *A dictionary of phonetics and phonology*. Routledge, 2004.

[33] Dimitrios Ververidis and Constantine Kotropoulos. Emotional speech recognition: Resources, features, and methods. *Speech communication*, 48(9):1162–1181, 2006.

[34] Yongjin Wang and Ling Guan. Recognizing human emotional state from audiovisual signals*. *Multimedia, IEEE Transactions on*, 10(5):936–946, 2008.

[35] Carl E Williams and Kenneth N Stevens. Emotions and speech: Some acoustical correlates. *The Journal of the Acoustical Society of America*, 52(4B):1238–1250, 1972.

[36] Martin Wöllmer, Florian Eyben, Stephan Reiter, Björn Schuller, Cate Cox, Ellen Douglas-Cowie, and Roddy Cowie. Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies. In *INTERSPEECH*, volume 2008, pages 597–600, 2008.

[37] Chung-Hsien Wu, Jen-Chun Lin, and Wen-Li Wei. Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. *APSIPA Transactions on Signal and Information Processing*, 3:e12, 2014.

[38] Jiahong Yuan and Mark Liberman. Speaker identification on the scotus corpus. *Journal of the Acoustical Society of America*, 123(5):3878, 2008.